Paper presented at the Sixth Meeting of the International Working Group on Price Indices, Canberra, Australia, 2-6 April 2001.

Scanner Indexes for the CPI

David H. Richardson[1]

December 25, 2000

1. Introduction

The Consumer Price Index (CPI) is the nation's primary measure of the price change of consumer goods and services. To produce the CPI, Bureau of Labor Statistics (BLS) staff track the prices of a sample of consumer items in the various categories of consumer spending in stores and other retail outlets. In some of those categories, virtually all of the items have a manufacturer-supplied Universal Product Code (UPC) printed on the products to be read by scanners. [See McKaig, 1999] Retailers set prices by UPC for efficiency at the checkout and for inventory management. Consequently the retailers create computerized records by UPC of the prices and number of units they sold, records that are commonly called *scanner data*. Scanner data are also used in marketing research to track promotions and variations in packaging, product size, and pricing. In a report to the Senate Finance Committee, the Boskin Commission [1996] suggested that scanner data could be used in the CPI for additional commodity detail, as have The Conference Board [1999] and de Haan and Opperdoes [1997a,b]. Previous BLS work on scanner data includes Bradley et. al. [1997] and Reinsdorf [1997].

The data used in the basic indexes of most item strata are collected by BLS field economists, typically by personal visits to the store or other retail outlet. In retail outlets such as supermarkets, the collection is by observation of the prices on the shelves. Each of these prices is called a quote, whether a price was actually collected or whether an unsuccessful attempt was made. The outlets and items are chosen using a probability of selection proportional to sales (PPS) approach. The outlet sample for most strata is based on the Point of Purchase Survey (POPS), a telephone inquiry of where items are purchased by consumers in the urban U.S. The result is that the CPI tracks prices in the outlets where people actually shop. Within the sample outlets, the item samples are selected in proportion to the outlets' sales within the item category. The same items are priced each month to the extent possible, and new items are substituted as the current items disappear from the market.[2] The outlet sample is rotated approximately every four years and new item samples are initiated at that time.

It has long been thought that scanner data can be of benefit to the Consumer Price Index for some or all of the following reasons:

---

[1] Many thanks to all of the members of the ScanData team; this paper is really a group effort. Special thanks to Ralph Bradley, Bill Cook, Dennis Fixler, John Greenlees, Bill Hawkes, and Walter Lane for helpful comments, to Lyubov Rozental for the programming assistance that made it all possible, and to Scott Pinkerton for Cereal analysis. All remaining errors are of course my own.

[2] Occasionally new outlets are chosen to augment the sample in cases where there has been serious attrition due to the closing of outlets.

1. Scanner indexes promise greater precision or lower variance.
2. Scanner data is a record of transactions that have actually taken place while the CPI collects the prices of items on the shelf whether transactions actually took place at these prices in the given month or not.
3. Scanner data results in indexes with finer "granularity"—greater commodity detail.
4. The scanner sample is more representative of the universe because the weights are estimated more accurately and updated more frequently.
5. The scanner data are cleaned according to rules that can be applied consistently and studied academically.
6. Scanner data presents the opportunity to implement superlative indexes.

This paper explores the ways that the CPI can use scanner data in place of traditionally collected price data.

The A.C. Nielsen Co. and Information Resources, Inc. are the only major U.S. vendors of supermarket scanner data, routinely collecting scanner data from the retailers, e.g., Kroger, Safeway, and Giant, and marketing it to the manufacturers, e.g., General Mills, Nabisco, and Pepsi. While the manufacturers could collect this data themselves, there are clear economies of scale enjoyed by the vendors, and in fact no manufacturer collects this data directly. The vendors add value by sorting the data into useful categories, e.g., cereal, subjecting them to range and consistency checks, and making them available in a standard format. The data contain three principal dimensions—product, time, and geography—and any desired subset is available. Thus BLS can theoretically request data for the items that are currently priced, i.e., according to BLS item definitions, and also for the CPI index areas. This paper is a progress report on a major CPI program initiative to construct scanner-based test indexes for Breakfast Cereal in the New York metropolitan area.

The CPI is calculated in two stages based on the BLS partition of the consumption universe into 211 item strata and 38 geographical index areas. In the first stage a basic index is calculated for each of the 211 x 38 = 8018 index area-stratum combinations. The basic indexes for most item strata are constructed using a weighted geometric mean (Geomean) formula; the few remaining item strata use a modified Laspeyres formula. The second stage of the calculation uses the Laspeyres formula to combine appropriate sets of basic indexes to yield various higher level aggregates. These aggregates include the national All-Items CPI along with intermediate-level aggregations, e.g., national Food and Beverages, or geographical areas such New York All-Items. The CPI for the New York Consolidated Metropolitan Statistical Area (referred to in the CPI as A101) is the aggregation of three basic index areas, A109 (New York City), A110 (New York-Connecticut Suburbs), and A111 (New Jersey-Pennsylvania Suburbs).

To evaluate the possibility of improving the accuracy of the CPI by using scanner data, BLS created the ScanData team. ScanData's objective was to determine whether it is feasible to incorporate scanner data into the monthly CPI production. The method has been to produce demonstration or test indexes for Breakfast Cereal in New York using scanner data. A success in producing such indexes would confirm that it is indeed possible to improve the CPI using scanner data, while a failure would be evidence to the contrary. The goal is indexes that:

1. Are produced on the CPI production schedule;
2. Cover the entire domain of a basic item/area stratum by combining or "amalgamating" scanner and CPI data to eliminate outlet and geographic gaps;
3. Are consistent with CPI sampling principles;
4. Are based on a sample that is rotated and refreshed at least as often as under the current CPI procedure;
5. Use both the standard CPI geometric mean formula and a superlative index formula;
6. Use data cleaned at least to current BLS standards;
7. Use prices with tax.

ScanData has constructed test indexes based on Nielsen Breakfast Cereal data for A101 New York monthly in real time since March 1998, and has similar indexes based on back data in a somewhat different format from September 1994 to January 1998. These data have been used to construct Geomeans, Laspeyres, Tornqvist, and Sato-Vartia indexes. The fixed weight indexes, i.e., Geomeans and Laspeyres, use weights based on the previous calendar year, while the weights for the Tornqvist and the Sato-Vartia are updated monthly. Soon cereal data will begin to flow for the whole Northeast census region, and beginning in January 2001, for the whole country.

The results show that, over the whole sample period, the scanner Geomean index was 104.9 while the CPI was 101.1 on a February 1998 = 100 base. On average, the scanner indexes have less variability than the current CPI. There are about 80,000 scanner quotes collected each month in New York and about 55 traditional CPI quotes, and hence there is a potential reduction in the standard error by a factor of about $\sqrt{80{,}000/55} = 38.1$. One indication of the reduction in the standard error is the mean absolute percent error (MAPE) about the mean. If the level were constant from month to month, the MAPE would be a measure of the spread of the distribution. As it is, the MAPE includes both the percentage error and the change in the level. Through June 2000, the CPI Geomean relatives have a MAPE of 2.36% compared with 1.14% for the scanner Geomean relatives, an indication that the scanner relatives are more precise.

Section 2 below discusses the majority of the known technical issues in the implementation of the scanner indexes, saving the unit values issue to Section 3. The scanner data are used to compute a variety of indexes, the formulae for which are presented in Section 4. Section 5 presents a sample of the results of the real time experiment from March 1998 to June 2000, and also earlier data referring to October 1994 through January 1998.

2. Technical Issues

The production of scanner indexes has involved dealing with a host of technical issues, some of which have been considered in the academic literature, while others have not. Most technical issues other than the index calculation are considered here, while the calculation itself is discussed in Section 4. Three technical issues involving unit values—item definition, weekly unit values, and organization[3] level as opposed to outlet level indexes—are defined and discussed in Section 3. A summary of the status of all these technical issues is found in Table 2.1, followed by a definition and a more detailed discussion. A 'Y' in the 'Envisioned' column indicates that ScanData has a solution to the issue in mind, while an 'N' is an indication simply that there is work to do. The solution to a given issue is said to be 'Designed' if there is a mathematical solution on paper or an agreement with some required outside party on the appropriate procedure. A solution is said to be 'Implemented' if it exists at present in the ScanData computer program. There are 'Results' if data have been produced using the implemented solution.

Table 2.1. Status of Scanner Issues

Issue	Envisioned	Designed	Implemented	Results
Quote Timing	Y	Y	Y	Y
Refreshing Sample	Y	Y	Y	Y
Quote Eligibility	Y	Y	Y	Y
Sales Taxes	Y	N	N	N
Amalgamation	Y	Y	Y	Y
Migrating Quotes	Y	N	N	N
Data Cleaning	Y	Y	Y	Y
Imputation	Y	Y*	Y*	Y*
Variances	Y	Y	Y	N
Geography	Y	Y	Y	N
Average Prices	Y	N	N	N
Supply Disruptions	Y	Y	Y	Y
Aggregation	Y	Y	Y	Y

* Decision being reconsidered.

Quote Timing

The CPI collects data for the first 18 business days (i.e., Mondays through Fridays excluding holidays) in the month, except 15 days in November and December. Each CPI

[3] The term "organization" is used rather than "chain" to denote what is normally referred to as a supermarket chain in order to avoid confusion with the technical term "index chaining." In addition some organizations which appear to be supermarket chains to the consumer actually consist of independently owned stores with a common logo.

quote is assigned to one of three pricing periods of five or six business days each, and all of the data are used to produce a monthly index. Nielsen scanner data, by contrast, are collected weekly (a week is Sunday to Saturday) and ScanData receives monthly shipments of data from Nielsen with either four or five weeks of data. The last week in each shipment is the third week of the calendar month, and the data are due ten business days after the end of the third week. The data have almost always arrived two or three days early. The first week of the month is defined as the first week with at least five days in the given month, irrespective of any holidays. Thus, the designated "first" week of the month occasionally includes the last day or two of the previous month, but never more than this. Since the prices generally refer to a whole week, they also are the prices for the actual first week of the month. ScanData is able to use data for the first three weeks of the month in the test indexes and can produce the test indexes a few days earlier than they would be needed for the regular CPI computation. The indexes compare prices in the first three weeks of the current month to the first three weeks of the previous month.

The result of these collection rules has been for calendar 2000 that the median CPI collection period covers the first through the 25th, while the median scanner collection period covers the second through the 22nd. Therefore

1. The CPI and the Scanner indexes cover very similar time periods within the month;
2. The CPI data are collected a little later in the month;
3. The scanner data explicitly cover Saturday, Sunday, and holidays.

Refreshing the Sample

As noted above, the CPI outlet and item samples are rotated (the old sample is dropped and a new one takes its place) every four years in current practice. Nielsen refreshes (adds units to) the outlet sample periodically to maintain sample size and to ensure that it continues to reflect the market in terms of the distribution of outlets by geography, organization, format type, size, and age. Since the scanner data consist of a census of the items in the category in the sampled stores, the scanner item samples are refreshed continuously. For the test indexes that use the Geomeans or the Laspeyres formulas, ScanData refreshes both the weights and the item sample each year using the expenditure patterns of the previous year. Thus the 1999 scanner weights for New York Cereal are based on 1998 expenditures, the weights for 2000 are based on 1999, and so on.[4]

Quote Eligibility

Quotes in the CPI are eligible for pricing if they were selected in the most recent PPS sample. Scanner quotes are eligible simply if the items have been sold in the previous year. Once a quote is selected, the CPI collects its price if it is present on the

[4] One exception should be noted: the Nielsen data for the three New York index areas became available in the current format only in February 1998, and hence both the 1998 and 1999 weights for New York were based on 11 months of data rather than the desired entire year. In addition the 1998 indexes were based on the 1998 weights rather than, as would have been preferred, the 1997 weights.

shelf regardless of whether it has been sold at that price recently or at all. Scanner prices, on the other hand, are "transaction" prices, i.e., there is a price if and only if the item is sold during the given week, regardless of whether it is on the shelf at any particular time. CPI quotes are imputed if they are unavailable for pricing when the CPI data collector appears. Scanner quotes are imputed if they are eligible and if there were no sales in the first three weeks.

Sales Taxes

For cereal as for most other items, the CPI collects prices net of any applicable tax. The sales tax is applied subsequently, using a secondary source that has the sales tax rates for all of the jurisdictions in which there are CPI outlets. Scanner prices are also collected untaxed. For confidentiality however, the vendors will not disclose the exact location of the outlets, so ScanData cannot add the sales taxes in the same manner as for CPI data. This has not been a problem heretofore since breakfast cereal is not taxed in the New York CMSA. However, it will become a problem as additional index areas are added. ScanData has not designed a solution to this problem as yet. It is hoped that Nielsen can help with the taxes, with the caveat that Nielsen may not realize what a complex problem it is. A second best solution, which BLS can implement independently, is to calculate a population-weighted average sales tax each month for each item stratum in each index area, and apply the average sales tax to all of the outlets in the index area.

Amalgamation

The current CPI is designed so that every item in every outlet in all of the urban areas of the United States has a chance of selection in accordance with the PPS methodology. The scanner database is a proper subset of the universe covered by the current CPI. Therefore the scanner index will not cover the appropriate universe unless the outlets not covered by scanner data, e.g., because of their type, e.g., independent grocers, or because of their location, are covered by retaining existing CPI quotes.[5]

The indexes reported in this paper combine preliminary scanner indexes (using scanner data only) with indexes constructed from CPI data for the part of the universe scanner leaves out, a process called amalgamation. In particular, from 1998 through 2000 in New York there were four CPI quotes not in the scanner universe, all of which were from mass merchandisers or wholesale clubs. The mass merchandisers will be covered in future Nielsen data deliveries. The results show that the amalgamation in the New York cereal indexes did make a difference, but this small sample does not provide much evidence as to whether it adds value.

At present the amalgamation uses CPI weights, which come from the POPS. For example, if the current CPI had ten quotes and nine of them were from outlets covered by scanner data, the price relative computed from the scanner data would receive the sum of the weights of the nine quotes, while the non-scanner quote would retain its original

---

[5] Since the scanner markets are based on television markets, they do not always contain the corresponding Census index areas as proper subsets. On the other hand, Nielsen has done quite a good job of reconfiguring their data to the CPI areas.

weight. To date the Geomean formula has been used for amalgamation. Since the Tornqvist and the Sato-Vartia indexes rely on current period quantities, these formulas cannot be used for amalgamation. It is intended to use Laspeyres amalgamation in the Laspeyres strata.

Migrating Quotes

The objective for a basic index of the CPI is to track changes in the prices paid by the residents of a particular index area for items in a particular item stratum. The CPI outlet surveys reveal that consumers make some purchases in outlets located outside the consumers' home areas. Hence some of the CPI quotes are not collected in the area to which the index refers, but represent purchases of consumers from the given area in other places because of travel, mail order, or other factors. Scanner data refer to the prices charged in an index area, which is not quite the same as the prices paid by the residents of the index area. Thus, the uncorrected scanner index will not represent the CPI objective. None of the current New York CPI cereal quotes has "migrated" out of the New York metropolitan area, although some of the current CPI quotes for index area A109 are collected in A110 and A111. For this reason, a more adequate treatment of migrating quotes would involve making the index for A109 a weighted average of the indexes for all three index areas. A more complete solution to this problem awaits the mapping of the whole country, in order to design a comprehensive solution.

Data Cleaning

Before indexes can be computed the data quality must be assessed and questionable observations deleted. This is an automatic process that corresponds to the current CPI procedure wherein a BLS economist (called a commodity analyst) investigates quotes with large price changes and makes decisions whether or not to use them in the index. While Nielsen does implement quality checks, for purposes of the CPI they have to be supplemented by BLS efforts. Therefore ScanData has developed a procedure for automatically cleaning the Cereal data based on the following five rules:

1. Accept all quotes that do not decline more than 37.5% nor increase more than 60% in a given month.
2. On quotes with a promotion on the lower price, accept all quotes that do not decline more than 60% nor increase more than 150%.
3. Accept all quotes where the elasticity implied by the two months is at least as large as the 1.0 implied by the Geomeans calculation. In other words, if the price goes down the volume must increase more than proportionately (so that revenue does not decline), and conversely.
4. Accept all quotes if, in the last 12 months, the price has been as high or higher than the current price, and if the price has been as low or lower than the current price.
5. Do not use any quote that does not satisfy at least one of 1.-4.

If the item is sold in multiple units, e.g., as a "2-fer" or a "3-fer," in one but not both of the relevant time periods, rules 1. and 2. are modified as follows:

1′. Accept all quotes that do not decline more than 50% nor increase more than 100% in a given month.

2′. On quotes with a promotion on the lower price, accept all quotes that do not decline more than 68% nor increase more than 212.5%.

Rule 5 applies in both cases, i.e., a quote must pass only one of rules <1 or 1′>, <2 or 2′>, or 3, or 4. It has been found that the application of these rules deletes a few clearly questionable quotes but accepts most of the data.

Imputation of Missing and Suspect Prices

In both the current CPI and the scanner index, the cleaned database contains a record for each quote. The record includes (if available) the quote's collected price, its effective (per ounce) price, and its derived price. The derived price (also per ounce) is either the effective price where acceptable or the imputed price otherwise. As in the CPI, only quotes with both an acceptable price in the current month and a derived price in the previous month are used in scanner index calculation.

Missing data for the scanner indexes (as for CPI data) are imputed for use in future months, and unacceptable prices are treated as missing data. In this process, missing prices are imputed by moving the last acceptable price forward by the (chain of) stratum relative(s). This is equivalent to current CPI practice, where the last acceptable price is moved forward each month by the index area-stratum relative. Once the indexes have been calculated, the imputed prices are calculated explicitly and entered as derived prices in the database for use in future months.

In the CPI, if an item is not on the shelf, is not expected to be restocked shortly, and has not been sold in the last few days, it is reported as missing. In cases where the item is not expected to return, the CPI Field Representative will select a replacement. In the scanner procedure, if an item were not sold in any of the first three weeks of the month, it would be treated as missing. In scanner data indexes however, replacements for missing items are not sought, since the data already includes all possible items.

There is a new issue that arises in full force with scanner data. Because of the 100% sampling rate for items within an outlet, new items appear in scanner data much more often than in the CPI; moreover there is virtually no lag between the time a new item appears in the outlet and the time it is available for our use. Often, at least for cereal, the prices of goods newly introduced into the market are not equilibrium prices but rather test prices set by a product manager in the hope of obtaining a relatively stable volume at an acceptable price. They are erratic, sometimes beginning quite low and then increasing to a level comparable to other similar products, while at other times beginning at a comparable level, declining in a deep sale, and then returning to the comparable level. This process can be repeated for a single item over its first few months. It can be fascinating to observe the resulting price trajectories, but any information on inflation is swamped by the wide variations in the prices resulting from the marketing process. The

CPI has long faced the similar problem of products that disappear after a dramatic price reduction or a "close out special." Since in formulas with constant weights (such as the Geomeans) the long run effect on an index of a product that appears and then, after some time disappears, is just the ratio of the first and last prices, the introduction and withdrawal of items from the market can have a nontrivial effect on the index. The solution to this problem is in an imputation system that is based on something other than the last acceptable collected price. Thus in Table 2.1 the imputation design is labeled as a decision being reconsidered.

Variances and Replicates

The CPI variances are computed from replicates[6]. In the current CPI the small sample sizes mean that there are just two replicates for most index areas to serve as the basis of an estimate of the sampling variance. The much larger scanner sample potentially could support many more replicates and hence a much more precise estimate of the variance. It has been proposed in the Scanner Project to use a stratified jackknife calculation based on a segmentation of the scanner sample into separate index area-organization-identified strata: one for each of 3-8 major organizations and one for the remaining scanner outlets within each index area. Within each stratum there would be clusters consisting of from one to three outlets each, and the replicates consist of the sample with each cluster deleted in turn. At present, however, scanner variances are a work in progress.

Northeast and National Geography

The population target of the CPI is the non-institutional population living in metropolitan or urban non-metropolitan areas. As noted the CPI has partitioned the urban United States into 38 index areas.[7] Scanner data are available for "markets," which are generally smaller than the U.S. Census-defined metropolitan areas. Nielsen is currently engaged in mapping the entire U.S. geography into the CPI index areas. There will be geographic gaps when this is complete since the 31 self-representing metropolitan areas consist of the 29 largest ones plus Honolulu and Anchorage, and Nielsen does not collect data for Anchorage.

Average Prices

The CPI computes average prices for a number of items for the convenience of the users. It is clear that extremely accurate average prices could be computed using scanner data. To date there has been no decision as to which average prices to compute, i.e., to simply continue the current set or expand it.

---

[6] See Swanson [1999]. A replicate is a sample of only part of the universe of available quotes. An estimate of the sampling variance can be obtained by comparing the indexes estimated from the replicates.

[7] The index areas are 31 self-representing metropolitan areas and 7 "region-size classes" for the remainder of the covered population; for details see Williams [1996].

Protection against Data Supply Disruptions

The CPI program collects most of the prices it uses; consequently, the data supply is controlled within the program, and the program is responsible for any data supply disruptions. Delegating the basic data collection to a scanner data vendor, however, creates the possibility of data supply disruptions that are not under the direct control of the CPI program. Therefore, in order for the program to fulfill its responsibility for a continuous flow of data, there must be a backup. One backup could be to continue to process the POPS as it is currently done to obtain:

1. The weights required for the scanner amalgamation procedure;
2. A sample of quotes that could be initiated as a fallback in case of need, including a vendor supply disruption.

No supply disruptions have occurred while we have been purchasing scanner data on a flow basis, and none is expected. The effect of a backup would be to preserve the ability of BLS to begin collecting prices for the scanner items in case of need.

Aggregation

ScanData is committed to building up higher level area aggregates for cereal as appropriate. It has been decided that, as additional areas are added, the scanner indexes will be combined using the current CPI aggregation weights and the current Laspeyres calculation.

3. Unit Value Issues

This section considers three technical issues: weekly unit values, item definition, and organization-level as opposed to outlet-level indexes. The reason for considering these issues together in a separate section is that they all touch on the unit value controversy in one way or another. A unit value is a quantity-weighted average price of an item. One way to compute a unit value is to divide the revenue for the item by the number of units sold. Unit values are used for the basic area-item indexes in most countries; they are not used in the United States because a weight is assigned to each individual quote.

One encounters unit values frequently in scanner data since, for example, if an outlet has two different prices of an item in a week, the reported price is often the revenue for the item divided by the number of units sold. On the other hand it may be appropriate to use unit values to combine observations over several weeks in the same month, or over outlets, or over similar items. Since a unit value is not a price in that often no one pays the unit value exactly, there is some controversy over whether our calculations should use them as if they were prices. ScanData is using unit values to combine the weeks of the month, UPC codes with minor differences, and the outlets within an organization in a given CPI area.

Weekly Unit Values

ScanData combines data for the three weeks of the month using unit values: the quantity-weighted average of prices for a given item.[8] The price relative of an item is its unit value (average over the three weeks) for the current month divided by its unit value for the previous month. ScanData then computes the Laspeyres, Geomean, and other price indexes using these price relatives of item unit values with the appropriate PPS weights. One condition for this to make sense is that the quantities all must be measured in the same units, a condition that is clearly satisfied here.

Instead of combining these prices and quantities in a unit value, one could combine them using the chosen index formula, in the same way that prices and quantities of different items are combined. The unit value approach more accurately reflects the preferences of the shopper who searches out the lowest prices each week, and also the consumer who stockpiles during a particularly good special, but then purchases nothing until the next special. [See Triplett, 1999] Consider the problems that arise by *not* using unit values and considering purchases in different weeks as different goods:

1. The weeks are arbitrarily defined, starting as early as the penultimate day of the preceding month, and as late as the fifth of the current month. Thus a purchase made every month on a particular date, say any day between the 6th and the 11th, will sometimes be allocated to the first week of the month, and sometimes to the second.
2. The commodities purchased in the different weeks seem to satisfy the same needs and desires on the part of the consumer.
3. Unit values are required at some level in order to construct an index at all.
4. Consumers who stockpile are not indexed correctly without unit values over the weeks.
5. *Not* using unit values implies an inherent rigidity in consumer behavior since it is assumed that the items in each of the three weeks are unrelated, and that the elasticity of substitution among them is zero.
6. Defining items with a finer granularity, as is the case if quotes in different weeks are treated as separate items, results in more missing data and more imputations.

Thus real inaccuracies can be introduced by *not* grouping identical commodities using unit values, and there is a powerful argument for considering purchases in the different weeks of the month as the same good.

Item Definition

Occasionally a manufacturer will keep the product constant but will create a new package with a new UPC code, a process called “churning.” Alternatively, and more interestingly, new UPC’s sometimes appear that involve only small changes in the package size or flavor, e.g., blueberry or raspberry, that (almost) always sell at the same price. Changes in the package size can be used to indirectly raise or lower the effective

---

[8] An item is commonly but not always a given UPC in a given store; see the subsection on *Item Definition* below.

(per ounce) price while keeping the shelf price constant. Both to reduce attrition and to capture these indirect price changes, it is important that these small to nonexistent variations in the product be grouped together into a single item for index calculation.

Each month Nielsen supplies a list of new UPC codes together with the respective sizes and product descriptions. These are compared to the UPC codes already in use and, if in the analyst's judgment the differences are sufficiently small that the products are interchangeable, they are combined into a single item. The quantity (in ounces) of these combined items is the sum of the quantities sold of the constituent UPC's, the expenditure is the sum of the expenditures, and the price is the average price or unit value per ounce.

Organization Level Indexes

The CPI is based on the price of a given item at a given outlet at a given time. Scanner data is available at the outlet level, and outlet data can easily be aggregated up to the organization level, in which case the quote is the unit value of a given item in a given organization at a given time. The organization level index is based on the unit value or average price of an item at the outlets in an organization.[9] It may be that unit values across outlets are appropriate: "if individual outlet data on transactions were not available or were considered to be too detailed, then unit values for a homogenous commodity over all outlets in a market area might form the lowest level of aggregation." [Diewert 1995, p. 22] In fact, "everyday life suggests—and the suggestion is confirmed by our coffee data set—that consumers easily switch between outlets in response to relative price changes." [de Haan 1999, p. 64] Thus it has been thought that aggregation across outlets is not particularly controversial, and aggregation across different outlets in the same organization would seem to be even less so. The extra detail provided by outlet-level data is not much trouble to process, but it turns out that organization-level data is less expensive than outlet-level data because there are only a fraction as many data points. Perhaps of more significance is the need to reflect adequately the shoppers who search out the best sale, on the one hand, as opposed to the ones who always shop at the same outlet.

So far, the organization level and outlet level indexes for cereal have been very similar. In a production mode, however, budget constraints may make organization level indexes appealing. Nevertheless, there are three compelling reasons to continue to receive outlet level data in a research project:

1. Outlet level data allow the study of the differences between organization level and outlet level indexes.
2. Outlet level data allow ScanData more control of the data quality.

---

[9] Each month the scanner data comes with codes for each organization, without identifying the particular organization. There is always an "Other" organization representing smaller stores and independents. Heretofore the organization level index has been computed treating the other stores as if they constituted an actual organization. It has been proposed to modify this by treating each of the other stores as if they were separate organizations.

3. Outlet level data facilitate variance estimation.

4. Index Formulae

The scanner project originally was intended to produce Geomean indexes comparable to the current CPI. However, the scanner sample leads naturally to superlative indexes since the quantity of each item is collected each month along with the price. Therefore, Tornqvist and Sato-Vartia formulas have also been computed, along with the Laspeyres. The status of the various alternative index calculations is summarized in Table 4.1 below.

Table 4.1. Status of Index Formulas

Index	Envisioned	Designed	Implemented	Results
Geomean	Y	Y	Y	Y
Monthly Tornqvist	Y	Y	Y	Y
Annual Tornqvist	Y	Y	N	N
Laspeyres	Y	Y	Y	Y
Sato-Vartia	Y	Y	Y	Y

As noted above the second stage of CPI calculation is aggregation of basic indexes. The CPI for New York cereal is the aggregate of three basic indexes. With respect to any index formula *X*, the New York CMSA or A101 index at *t* compared with base month *b* is calculated from what are referred to in CPI terminology as "cost weights." The month *t* cost weight for A101 is aggregated from the constituent index area cost weights,

$$(4.1) \quad C_X^t = \sum_{m=1}^{3} C_{Xm}^t .$$

The cost weight for index area *m* at time *t* is the product of A_m, the population-expenditure weight for cereal in index area *m*, and all of the period-to-period index relatives since the base time *b*,

$$(4.2) \quad C_{Xm}^t = A_m \prod_{p=b+1}^{t} R_{Xm}^{p,p-1} .$$

The index of the price change from *t* - *k* to *t* with respect to index formula *X* is just the ratio of the relevant cost weights from (4.1),

$$(4.3) \quad X^{t,t-k} = C_X^t / C_X^{t-k} .$$

The calculation (4.1) - (4.3) corresponds exactly to the current CPI.

Geomean

The calculation of the Geomean relative used in the current CPI is $R_{Gm}^{t,t-1} = \prod_{j} I_{j}^{t,t-1} \left(p_{j}^{t} / p_{j}^{t-1} \right)^{S_{j}^{0}}$. Here j indexes the quotes in index area m, $I_{j}^{t,t-1}$ is the indicator function for the presence of quote j at both months t and t - 1, and the quotes are weighted by S_{j}^{0}, the expenditure share of quote j at time 0, $S_{j}^{0} = E_{j}^{0} / \sum_{i} E_{i}^{0}$.

The scanner Geomean relative in index area m from month t - 1 to t is

$$(4.4) \quad R_{Gm}^{t,t-1} = \prod_{c=1}^{n_m} \prod_{j} I_{j}^{t,t-1} \left(p_{j}^{t} / p_{j}^{t-1} \right)^{S_{cj}^{0}} .$$

In (4.4), each index area m is composed of n_m organizations c, j indexes the different quotes within an organization, and the quotes are weighted by the expenditure share of quote j in organization c at time 0, except for 1998 the previous calendar year,

$$(4.5) \quad S_{cj}^{0} = E_{cj}^{0} / \sum_{i} E_{ci}^{0} .$$

The price p_{j}^{t} is the unit value of commodity j in organization c at time t,

$$(4.6) \quad p_{j}^{t} = \sum_{i=1}^{n_{mc}} P_{ij}^{t} Q_{ij}^{t} / q_{j}^{t} ,$$

where P_{ij}^{t} and Q_{ij}^{t} are the prices and quantities, respectively, of j at the different outlets at time t, $q_{j}^{t} = \sum_{i=1}^{n_{mc}} Q_{ij}^{t}$ and there are n_{mc} quotes in each organization c. The A101 cost weight, using (4.4) in (4.2) with $X = G$, and then substituting into (4.1) and (4.3), results in the Geomean index $G^{t,t-k} = 100 C_{G}^{t} / C_{G}^{t-k}$.

Monthly Tornqvist

This is a chained Tornqvist calculation, the geometric mean of the item relatives with weights equal to the average expenditure shares of the current and preceding months. Thus with $W_{cj} = \left(S_{cj}^{t-1} + S_{cj}^{t} \right) / 2$,

$$(4.7) \quad R_{Tm}^{t,t-1} = \prod_{c=1}^{n_m} \prod_{j} I_{j}^{t,t-1} \left(p_{j}^{t} / p_{j}^{t-1} \right)^{W_{cj}} .$$

The A101 cost weight, using (4.7) in (4.1) - (4.3) with $X = T$ results in the Tornqvist index $T^{t,t-k} = 100 C_{T}^{t} / C_{T}^{t-k}$. The monthly Tornqvist relative is a standard output and has been amalgamated from March 1998 forward. This amalgamation perforce uses the Geomean formula since there are no monthly expenditure weights for the CPI quotes.

Annual Tornqvist

In this form of the calculation, the base for the Tornqvist would be one fixed month that is used for a whole year. The current and base periods are single months. Say

January were chosen as the base month. In this case, the February relative would be calculated as in the monthly Tornqvist above. The March relative, however, would be the Tornqvist of the March to January prices as if the February data did not exist, i.e., both the time 0 prices and expenditure shares would refer to January. This would give up the pretense of the discrete time approximation to the Divisia, but would eliminate any monthly chaining bias.

Laspeyres Index

This is the textbook Laspeyres, without any correction for formula bias, using weights updated once per year just as the Geomean. Thus the Laspeyres relative is

$$(4.8) \qquad R_m^{t,t-1} = \frac{\sum_{c=1}^{n_m} \sum_j p_j^t q_j^{t-1}}{\sum_{c=1}^{n_m} \sum_j p_j^{t-1} q_j^{t-1}} = \sum_{c=1}^{n_m} \sum_j I_j^{t,t-1} S_{cj} \left(p_j^t \big/ p_j^{t-1} \right),$$

where the shares S_{cj} are defined in (4.5). The A101 cost weight, using (4.8) in (4.1) - (4.3) with $X = L$, results in the Laspeyres index $L^{t,t-k} = 100 C_L^t / C_L^{t-k}$. The Laspeyres provides an upper bound with which to can test the other indexes.

Sato-Vartia Index

The Sato-Vartia relative is a logarithmically weighted geometric average of the price relatives where the weights W_{cj} are proportional to $m_{cj} = \dfrac{S_{cj}^t - S_{cj}^{t-1}}{\ln(S_{cj}^t) - \ln(S_{cj}^{t-1})}$, so that

$$(4.9) \qquad R_{Sm}^{t,t-1} = \prod_{c=1}^{n_m} \prod_j I_j^{t,t-1} \left(p_j^t \big/ p_j^{t-1} \right)^{W_{cj}} .$$

The A101 cost weight, using (4.9) in (4.1) - (4.3) with $X = S$, results in the Sato-Vartia index $S^{t,t-k} = 100 C_S^t / C_S^{t-k}$. This amalgamation uses the Geomean formula since there are no expenditure weights for the CPI quotes. Along with the Fisher, the Sato-Vartia satisfies more of the statistical axioms describing a desirable index than any other does. However, like the Tornqvist, it does not satisfy the monotonicity axiom. [Reinsdorf and Dorfman, 1999]

5. Results

Scanner indexes have been calculated for the three New York index areas beginning with February 1998 = 100 and continuing through June 2000. Over this period, cereal inflation in the New York CMSA as measured by the A101 CPI was 101.1, an increase of 1.1% over 28 months.[10] As we see from Table 5.1 below, the scanner indexes, both amalgamated and not amalgamated, showed quite a bit more inflation than the CPI. Unless there is some systematic divergence in price change between the scanner

[10] The BLS does not publish an index for breakfast cereal below the national level. The New York cereal index levels reported here are unpublished, due largely to the small sample sizes on which they are based and the associated high variances. Note also that the CPI switched from the Laspeyres to the Geomean for the lower level indexes in January 1999.

universe and the CPI universe, the scanner indexes will give a more accurate measure of inflation. However, since the four CPI quotes that since September 1999 have been amalgamated come from one mass merchandiser and one wholesale club, there may be such a systematic divergence, at least in the short run. In addition, there was a sharp dip in the A101 CPI in October 1999 for which there is no apparent explanation. The A101 CPI had only partly recovered by June 2000, the end of the sample period. Looking ahead, this pattern appeared in a much attenuated form, if at all, in the national cereal CPI. The scanner indexes increased sharply in October 1999, but this mainly served to reverse an equally sharp decrease in September.

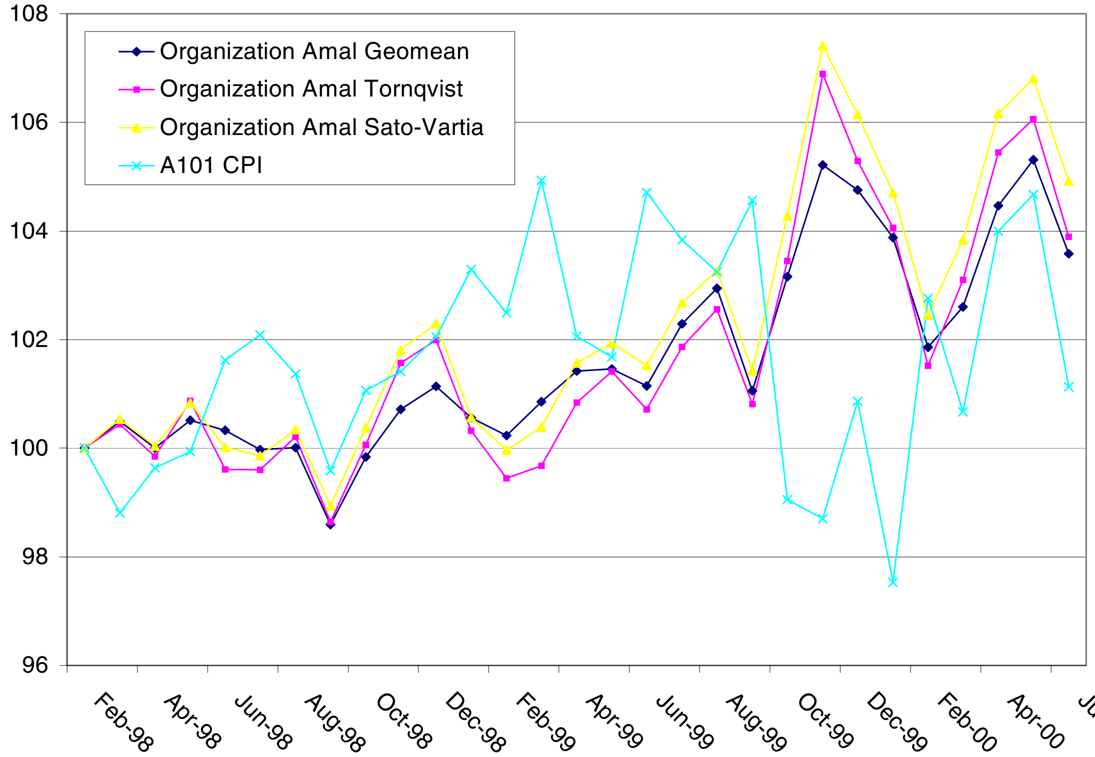
Over the whole 28-month sample period, the A101 CPI was 2.5% lower than the amalgamated Geomean index and 3.8% lower than the scanner Geomean. In addition, the Tornqvist, in both its amalgamated and scanner forms, showed more inflation than the Geomean, and the Sato-Vartia showed more still. The fact that the Tornqvist showed more inflation than the Geomean is evidence that, at least for New York cereal, the elasticity of substitution is less than the 1.0 assumed by the Geomean.

Table 5.1. Scanner Indexes, 1998 - 2000

Month	A101 CPI	Geomean		Tornqvist		Sato-Vartia	
		Scanner	Amalg.	Scanner	Amalg.	Scanner	Amalg.
Feb-98	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Mar-98	98.8	100.5	100.5	100.4	100.4	100.5	100.5
Apr-98	99.6	100.0	100.0	99.8	99.8	100.0	100.0
May-98	99.9	100.6	100.5	100.9	100.9	100.9	100.8
Jun-98	101.6	100.3	100.3	99.6	99.6	100.0	100.0
Jul-98	102.1	100.0	100.0	99.6	99.6	99.9	99.9
Aug-98	101.4	100.5	100.0	100.7	100.2	100.8	100.3
Sep-98	99.6	98.2	98.6	98.2	98.6	98.6	98.9
Oct-98	101.1	99.3	99.8	99.6	100.1	99.9	100.4
Nov-98	101.4	101.0	100.7	101.9	101.6	102.2	101.8
Dec-98	102.1	101.5	101.1	102.4	102.0	102.7	102.3
Jan-99	103.3	101.0	100.6	100.7	100.3	101.0	100.6
Feb-99	102.5	100.8	100.2	99.9	99.4	100.5	100.0
Mar-99	104.9	100.9	100.9	99.7	99.7	100.4	100.4
Apr-99	102.1	101.4	101.4	100.7	100.8	101.5	101.6
May-99	101.7	102.8	101.5	102.8	101.4	103.3	101.9
Jun-99	104.7	102.5	101.1	102.1	100.7	102.9	101.5
Jul-99	103.8	103.7	102.3	103.3	101.9	104.1	102.7
Aug-99	103.2	103.5	102.9	103.1	102.6	103.8	103.3
Sep-99	104.6	101.1	101.1	100.8	100.8	101.5	101.4
Oct-99	99.1	103.5	103.2	103.8	103.4	104.7	104.3
Nov-99	98.7	105.6	105.2	107.4	106.9	108.0	107.4
Dec-99	100.9	105.1	104.8	105.7	105.3	106.6	106.1
Jan-00	97.5	103.4	103.9	103.6	104.1	104.3	104.7
Feb-00	102.8	102.5	101.9	102.1	101.5	103.1	102.5
Mar-00	100.7	103.4	102.6	103.9	103.1	104.7	103.8
Apr-00	104.0	105.3	104.5	106.4	105.4	107.2	106.2
May-00	104.7	105.6	105.3	106.4	106.1	107.2	106.8
Jun-00	101.1	104.9	103.6	105.3	103.9	106.3	104.9

The amalgamated indexes from Table 5.1 are plotted against the CPI in Fig. 5.1. The CPI used the Laspeyres formula, an upper bound to the true cost of living index, through December 1998 and the A101 CPI shows more inflation than the scanner indexes through September 1999 before dropping quite a bit below. The higher level of measured inflation in the scanner indexes as opposed to the CPI is apparent at the end of the sample period in June 2000, despite the Laspeyres calculation through December 1998.

Figure 5.1. Amalgamated Indexes and the CPI, 1998 - 2000



The corresponding scanner indexes, i.e., with no amalgamation, are plotted in Fig. 5.2 below. While there is no particular reason to expect that the Geomean, Tornqvist, and Sato-Vartia should bear any particular relationship to each other, here the superlative Tornqvist and the almost superlative Sato-Vartia track each other extremely closely through January 1999, but then they diverge so that at the end of the period the Tornqvist is quite a bit below and is actually closer to the Geomean.[11] The Geomean calculation is more often below the Tornqvist, and over the whole 28 months is also lower. By itself, this would be evidence that the elasticity of substitution is less than the 1.0 assumed by the Geomean. However, while this difference is clear over the whole sample, it is not particularly apparent in the month to month data, since the Geomean shows a greater percent change than the Tornqvist in 13 of the 28 months. Similarly, while the Sato-Vartia is generally above the Tornqvist, the latter shows a greater percent change in 15 of the months.

[11] The mean absolute difference (MAD) between the Tornqvist and the Geomean was 0.46% while the MAD between the Tornqvist and the Sato-Vartia was only .16%.

Figure 5.2. Scanner Indexes and the CPI, 1998 - 2000

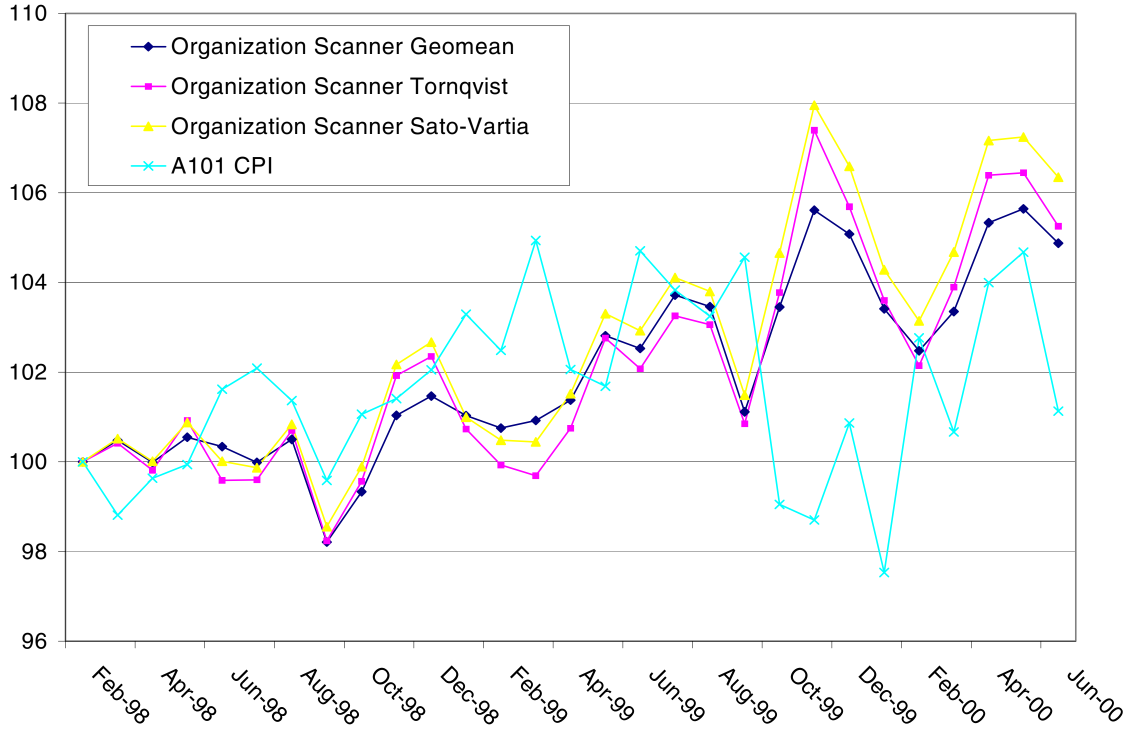
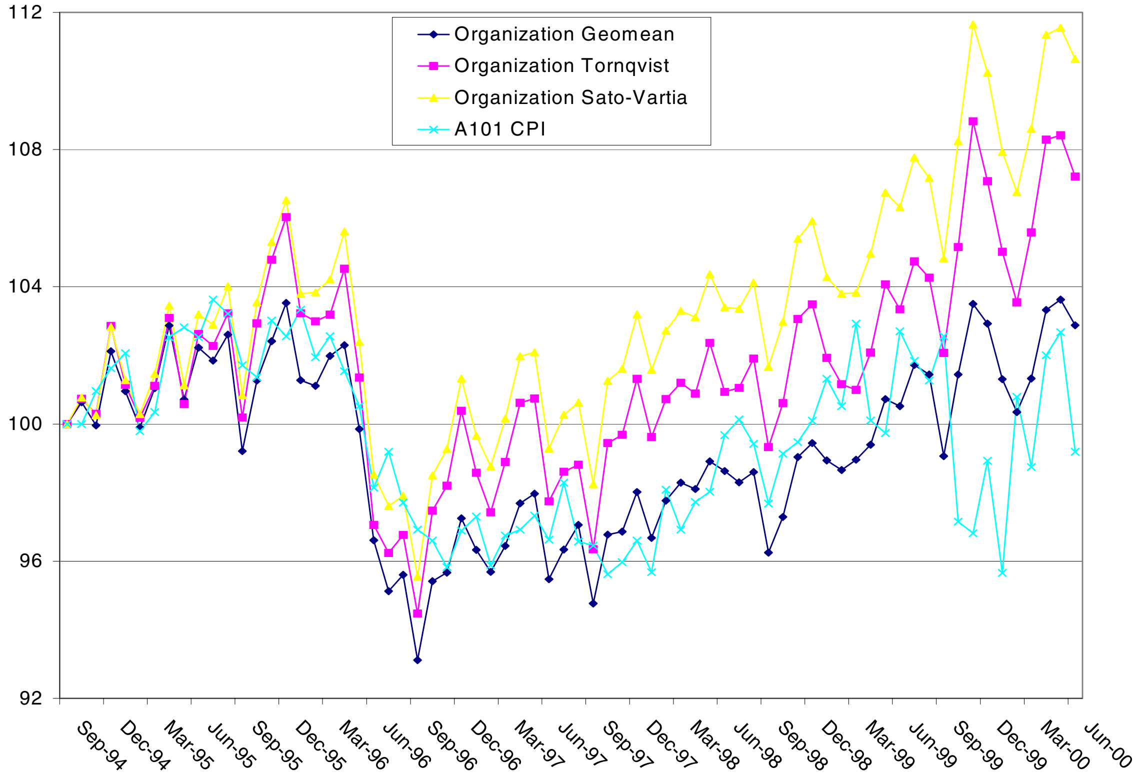


Fig. 5.3 below considers the scanner indexes without amalgamation back to September 1994, the beginning of the cereal data available to ScanData.[12] Here the final ranking of the formulas is the same as in the more recent period, but the Tornqvist is now closer to the Sato-Vartia over most of the sample period, and also at the end.[13] Despite using the Laspeyres formula through December 1998, the A101 CPI shows less inflation than the Tornqvist and the Sato-Vartia over the whole period, although it was above the Tornqvist as late as September 1999. The Geomean ends up quite a bit below the Tornqvist, providing further evidence that the elasticity of substitution is less than unity. Once again, while this difference is clear over the whole sample, it is not particularly apparent in the month to month data, since the Geomean shows a greater percent change than the Tornqvist in 38 of the 69 months. Similarly, while the Sato-Vartia is above the Tornqvist after March 1995, the latter shows a greater percent change in 39 of the months.

[12] Fig. 5.2 is subject to the caveat that the format of the scanner data changed between January and February 1998 in a way that caused a break in the series. This problem was resolved here by arbitrarily making all of the scanner relatives for February compared with January 1998 equal to the corresponding national CPI relative. Note also that (except for the A101 CPI) the data in Fig. 5.3 are the scanner indexes *not* amalgamated with any CPI data. The reason for this is that the data from September 1994 through January 1998 was not received in real time, and hence it was not possible to amalgamate the CPI quotes not covered in the scanner universe.

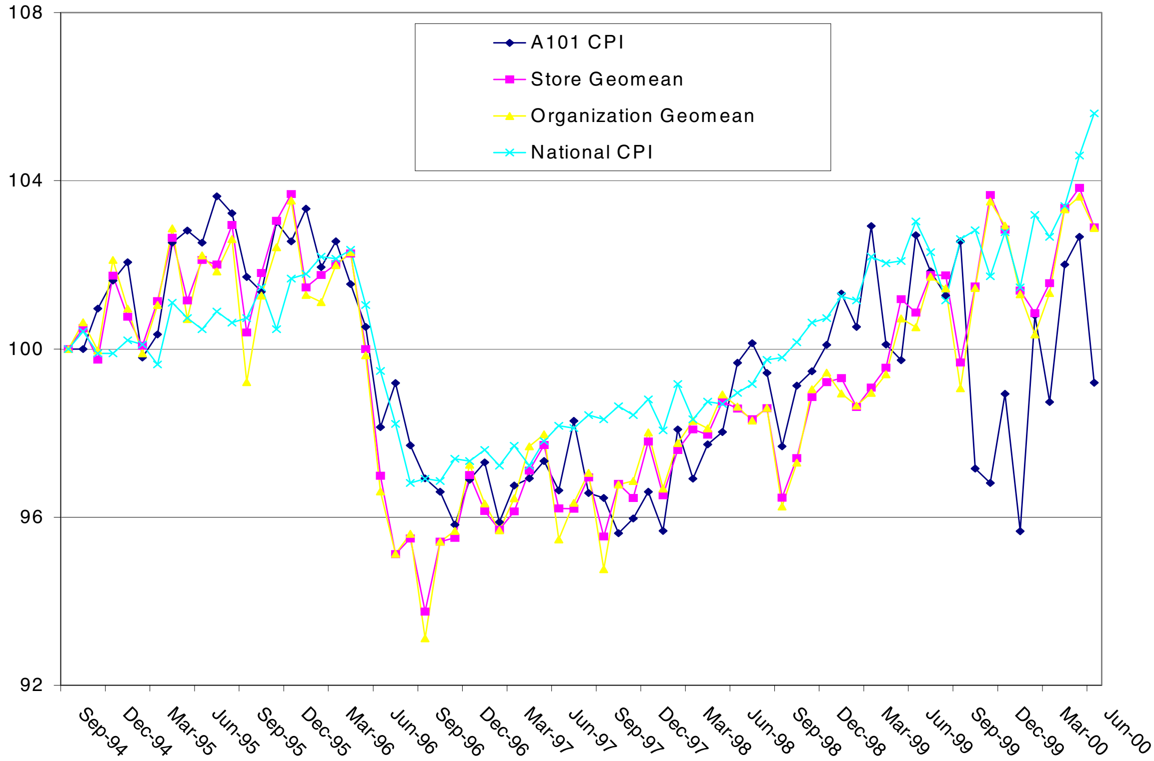
[13] From 1994 to 2000 the MAD between the Tornqvist and the Geomean was 0.45% while the MAD between the Tornqvist and the Sato-Vartia was only .13%, almost the same as the earlier period.

Figure 5.3. Scanner Indexes and the CPI, 1994 – 2000.



In Fig. 5.4 below we plot the national CPI for Breakfast Cereal against the A101 CPI and the scanner Geomean indexes at the outlet and organization levels. The organization level and outlet level Geomean indexes follow each other extremely closely and differ by less than 0.02% after 68 months, an indication that the unit value at the organization level doesn't make any difference. The national CPI and the A101 CPI do not track each other particularly closely, however, and hence, despite possible differences between the national and New York markets, we must suspect that there is considerable small sample variability in the A101 CPI. In particular the pronounced October 1999 dip in the A101 CPI is reflected only slightly in the national CPI. At the end of the study period the national CPI is above the New York scanner indexes while the A101 CPI is below them.

Figure 5.4. Scanner Geomean Indexes, A101 and National CPI, 1994 – 2000.



Returning to the 1998-2000 sample, the 1.1% increase in the CPI from Table 5.1 corresponds to an annual inflation rate of 0.5% over the 28 months from February 1998 to June 2000. Tables 5.2 below again shows that the annual scanner measured inflation rates were quite a bit higher than the CPI for any of the scanner indexes. As with all of these results, the interpretation must be very tentative because of the limited geography and product dimensions and the small CPI sample size.

Table 5.2. Full Sample Annual Inflation Rates

	Scanner	Amalgamated	Mean
CPI	---	---	0.5%
Geomean	2.1%	1.5%	1.8%
Tornqvist	2.2%	1.7%	1.9%
Sato-Vartia	2.7%	2.1%	2.4%
Mean	1.9%	1.4%	1.6%

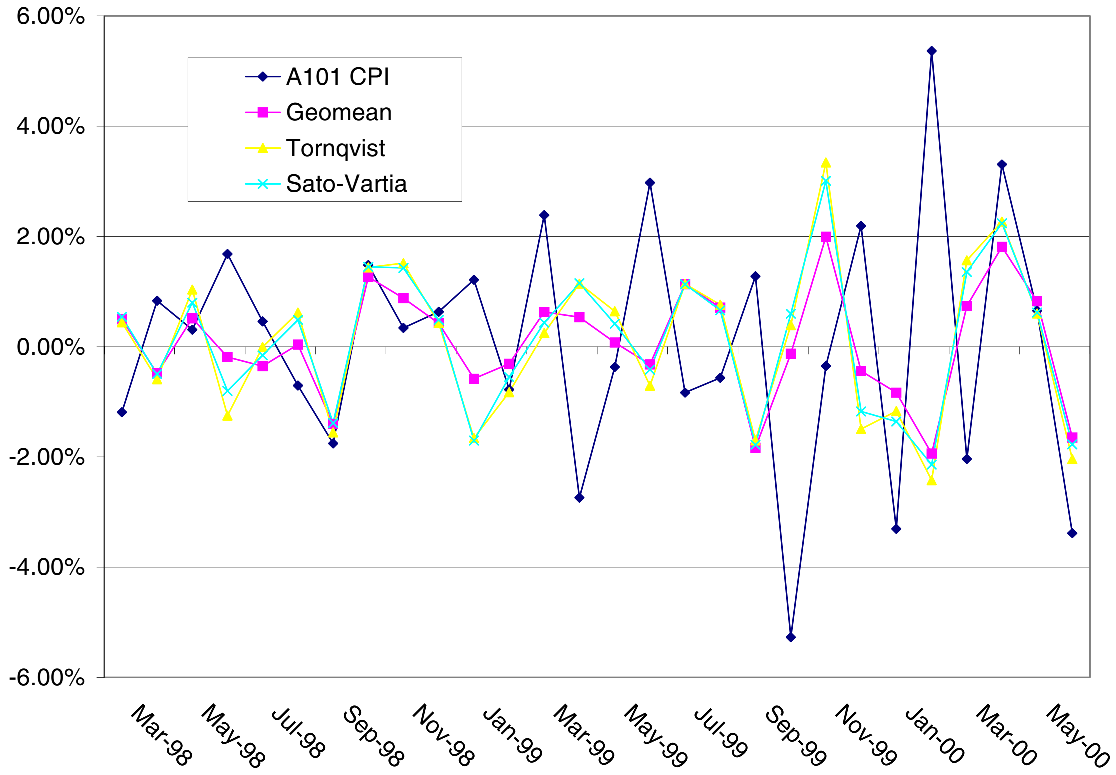
Remember that the MAPE, the mean percentage difference from the full sample mean, for the New York Cereal CPI relatives was 2.36%. This is much larger than the MAPE's for the different scanner calculations, as shown in Table 5.3, and hence the New York Cereal CPI is much more variable from month to month. The Geomean, whether scanner or amalgamated, wanders least of all.

Table 5.3. Full Sample MAPE's

	Scanner	Amalgamated	Mean
A101 CPI	---	---	2.36%
Geomean	1.09%	1.17%	1.13%
Tornqvist	1.49%	1.59%	1.54%
Sato-Vartia	1.40%	1.51%	1.45%
Mean	1.33%	1.42%	1.37%

The higher variability of the CPI and the lower variability of the Geomean are also shown in the following Fig. 5.5.

Figure 5.5. Monthly Percent Changes



References

Abraham, Katharine G., John S. Greenlees and Brent R. Moulton [1998] "Working to Improve the Consumer Price Index. *J. of Economic Perspectives*, pp. 27-36.

Balk, Bert M. [1999] "On the Use of Unit Value Indices as Consumer Price Subindices." in Walter Lane, ed., Proceedings *of the Fourth Meeting of the International Working Group on Price Indexes*. Washington DC, Bureau of Labor Statistics, pp. 112 - 120.

Boskin, Michael J., Ellen R. Dulberger, Robert J. Gordon, Zvi Griliches, and Dale Jorgenson [1996] *Final Report of the Commission to Study the Consumer Price Index*.

Bradley, Ralph, Bill Cook, Sylvia G. Leaver and Brent R. Moulton [1997] "An Overview of Research on Potential Uses of Scanner Data in the U.S. CPI." In *Proceedings of the Third Meeting of the International Working Group on Price Indices*, Bert M. Balk, ed. Voorburg: Statistics Netherlands.

Conference Board, The [1999] *Measuring Prices in a Dynamic Economy*.

Diewert, W. Erwin [1976] "Exact and Superlative Index Numbers." *J of Econometrics 4*, pp. 115-145.

----- [1983] "The Theory of the Cost of Living Index and the Measurement of Welfare Change." In *Price Level Measurement*, W. E. Diewert and C. Montmarquette (eds.) , Ottawa: Statistics Canada.

----- [1995] *Axiomatic and Economic Approaches to Elementary Price Indexes*. Univ. of British Columbia Dept. of Economics, Discussion Paper 95-01.

Feenstra, Robert C. and W. Erwin Diewert [2000] "Imputation and Price Indexes: Theory and Evidence from the International Price Program. Mimeo, U.S. Bureau of Labor Statistics.

Fixler, Dennis [1993] "The Consumer Price Index: Underlying Concepts and Caveats." *Monthly Labor Review* (December) pp. 3-12.

Greenlees, John S. [2000] "Consumer Price Indexes: Methods for Quality and Variety Change." *Statistical J. of the United Nations* 17, pp. 59-74.

Haan, Jan de [1999] *Empirical Studies on Consumer Price Index Construction*. Voorburg: Statistics Netherlands.

----- and Eddy Opperdoes [1997a] "Estimation of the Coffee Price Index Using Scanner Data: The Choice of the Micro Index. In *Proceedings of the Third Meeting of the International Working Group on Price Indices*, Bert M. Balk, ed. Voorburg: Statistics Netherlands.

----- and ----- [1997b] "Estimation of the Coffee Price Index Using Scanner Data: The Sampling of Commodities. In *Proceedings of the Third Meeting of the International Working Group on Price Indices*, Bert M. Balk, ed. Voorburg: Statistics Netherlands.

McKaig, S. Kate "Universal Product Codes," *CPI Tech Memo*, April 1999.

Moulton, Brent [1996] "Estimation of Elementary Indexes of the Consumer Price Index." *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

Reinsdorf, Marshall B. [1997] "Constructing Basic Component Indexes for the U.S. CPI from Scanner Data: A Test Using Data on Coffee." BLS Working Paper 277, April 1996.

----- and Alan H. Dorfman [1999] "The Sato-Vartia Index and the Monotonicity Axiom." *J. of Econometrics* 90, pp. 45-61.

Richter, Marcel [1966] "Invariance Axioms and Economic Indexes." *Econometrica 34*, pp. 739-755.

Swanson, David [1999] "Variance Estimates for Changes in the Consumer Price Index, January 1998-December 1998," *CPI Detailed Report* (December)

Szulc, Bohdan J. [1983] "Linking Price Index Numbers." In *Price Level Measurement*, W. E. Diewert and C. Montmarquette (eds.) , Ottawa: Statistics Canada, pp. 537-566.

Tornqvist, Leo [1936] "The Bank of Finland's Consumption Price Index." *Bank of Finland Monthly Bulletin* 10, pp. 1-10.

Triplett, Jack E. [1998] "Elementary Indexes for a Consumer Price Index. In *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, Walter .Lane, ed. Washington DC, United States Dept. of Labor.

----- [1999] "Should the Cost-of-Living Index Provide the Conceptual Framework for a Consumer Price Index?" Paper prepared for the fifth meeting of the International Working Group on Price Indices.

U.S. General Accounting Office [1999] *Consumer Price Index: Impact of Commodity Analysts' Decision Making Needs to be Assessed*. GAO/GGD-99-84.

Williams, Janet [1996] "The Redesign of the CPI Geographic Sample." *Monthly Labor Review* (December) pp. 10-17.